学校编码：10384 密级_____

学号：19020071152089

厦门大学

硕 士 学 位 论 文

分层线性模型在骨科住院患者医疗费用研究中的应用

Hierarchical Linear Models' Application in the Research of Medical Expenses of Orthopedics Inpatients

马 骅

指导教师姓名： 张志强 副教授

专 业 名 称： 应 用 数 学

论文提交日期： 2010 年 5 月

论文答辩日期： 2010 年 6 月

2010 年 5 月

厦门大学学位论文原创性声明

本人呈交的学位论文是本人在导师指导下,独立完成的研究成果。本人在论文写作中参考其他个人或集体已经发表的研究成果，均在文中以适当方式明确标明，并符合法律规范和《厦门大学研究生学术活动规范（试行）》。

另外,该学位论文为(　　)课题(组)的研究成果，获得（　　）课题（组）经费或实验室的资助，在（　　）实验室完成。（请在以上括号内填写课题或课题组负责人或实验室名称，未有此项声明内容的，可以不作特别声明。）

声明人（签名）：

年　　月　　日

厦门大学学位论文著作权使用声明

本人同意厦门大学根据《中华人民共和国学位条例暂行实施办法》等规定保留和使用此学位论文，并向主管部门或其指定机构送交学位论文（包括纸质版和电子版），允许学位论文进入厦门大学图书馆及其数据库被查阅、借阅。本人同意厦门大学将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于：

（　　）1.经厦门大学保密委员会审查核定的保密学位论文，于　　年　　月　　日解密，解密后适用上述授权。

（　　）2.不保密，适用上述授权。

（请在以上相应括号内打“√”或填上相应内容。保密学位论文应是已经厦门大学保密委员会审定过的学位论文，未经厦门大学保密委员会审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。）

声明人（签名）：

年　　月　　日

摘 要

分层线性模型由 Lindley 和 Smith 于 1972 年提出，国外主要用于社会科学、基础医学、临床医学、预防医学等领域的研究，国内多用于心理学领域。由于其可以处理层次结构的数据，非独立性资料，发掘深层次的数据结构等优点，使其可以解决传统模型难以解决的问题。

本文利用其研究在某一家骨科医院内住院病人的总医疗费用。每个患者的医疗总费用除了受个人的属性特征影响外，还受组织特征疾病种类的影响。我们将病种作为分层线性模型的组织层次变量，并将病种划分为骨折病种与非骨折病种进行分析，取得了非常好的效果。结论包括：

（1）病种之间在费用上存在显著差异。

（2）住院天数、治疗费用的差别效应在病种间存在显著差异。

（3）对单个病人而言，病人的住院天数、治疗费用与总费用显著正相关，住院天数和治疗费用可以解释总费用变异的约 94％。

（4）对所有病人而言住院病人总费用差异的大约 29％是病种差异引起的。

（5）对于平均总费用较高的病种，病人的住院天数差别效应小于那些平均总费用较低的病种，而病人的治疗费用差别效应大于那些平均总费用较低的病种。

（6）对病种而言，平均住院天数高的病种的天数差别效应较小。

（7）对病种而言，平均住院天数与病种平均费用正相关，平均住院天数与治疗费用差别效应正相关。

（8）对病种而言，平均费用变异中的 74％、天数差别效应变异中的 50％、治疗费用差别效应变异中的 46%能够用平均住院天数和是否骨折加以解释。

并且，在得到这些有效的结论之前，我们对模型进行了严格的检验。其中，利用怀特检验来检查组织内部的异方差性，是我们受到一般线性回归异方差检验的启示而来对分层模型进行尝试的。其他我们给出的实际检验的程序和操作方法也是反复实践和尝试的结果。

关键词：分层线性模型 模型检验 医疗费用 骨科医院 住院天数

Abstract

Hierarchical Linear Model was presented by Lindly and Smith in 1972. It is applied in the research field of social sciences, basic medicine, clinical medicine and preventive medicine by foreign countries and psychology by Chinese researchers. With advantages in processing and mining hierarchical structure or dependent data, many unsolvable problems under traditional methods could be answered.

This paper discusses its application in the study on medical expenses of orthopedics inpatients. The medical expenses of every orthopedics inpatient are not only influenced by individual characteristics, but also by disease categories. We take the disease categories as variables in organizational level and divide disease into facture and non-fracture. We find:

1. There are significant differences in medical expense between disease categories.

2. There are significant differences in the relationship of hospitalization day, treatment fee and total fee between disease categories.

3. The hospitalization day and treatment fee have positive correlation with total fee. The variables of hospitalization day and treatment fee can explain the 94% of total fee's variance.

4. 29% of the total fee differences between inpatients are caused by disease category.

5. For disease with high average total fee, the influence on total expense caused by hospitalization day is greater and the influence on total expense caused by treatment fee is smaller.

6. The disease with high average hospitalization day has relatively small differential effects.

7. Average expenses of diseases and average differential effects of treatment fee have positive correlation with average hospitalization day respectively.

8. 74% of the average expenses difference, 50% of differential effects difference of the hospitalization day, 46% of differential effects of the treatment fee are caused by average hospitalization day and fracture respectively.

Moreover, we have tested this model rigorously before having these effective conclusions. Under the inspiration of general linear regression, we use White's Test to test the heteroscedasticity in level-1. The other presented programs and operations on model test are tried by us repeatedly.

Key words: Hierarchical Linear Model; Model Test; Medical Expense; Orthopedics Hospital; Hospitalization Day.

目 录

Contents

第一章 问题的产生

1.1 回归模型的假设

回归分析（regression analysis）可以说是应用最广泛的统计分析技术，它主要是用来分析一群解释变量（或称自变量，independent variable）对被解释变量（或称因变量，dependent variable）的影响，不仅可以用在横断面的社会科学研究里，亦可以应用在纵贯面时间序列的分析上。在研究方法中，回归分析最常用在问卷调查的数据分析。除此之外，另一种研究模式——实验设计的数据分析，也可以通过回归分析技术得到相同的方差分析结果。

回归模型的应用非常广，它和一般线性模型（general linear model）一样，对于模型中的误差项假设或是因变量的假设相当严格。例如，以简单回归分析为例，

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

其中，自变量为X，因变量为Y，回归系数分别为β_0和β_1，而误差项为ε。回归分析对误差项的假设是：误差项彼此之间是独立，且服从平均数为0，方差为一常数的正态分布，即：

$$\varepsilon_i \sim iidN(0,\sigma^2)$$

有时我们是以对Y进行假设，给定X后的Y条件均值为$\beta_0 + \beta_1 X$，给定X后的Y条件方差为σ^2，则回归模型对Y的假设为：

$$E(Y_i|X_i) = \beta_0 + \beta_1 X_i,\ \ Var(Y_i|X_i) = \sigma^2$$

所以，在回归模型中，数据的特性特别是误差项或是因变量的独立性就很重要，因为它关系到我们使用的统计分析方法，是否符合统计理论中对模型的假设。所以正规数据分析的程序，会对统计模型分析后所产生的残差项进行残差分析（residual analysis），以检查数据分析后的残差是否符合统计模型对误差项的假设。

但是在一般社会横断面（cross-sectional）的研究中，如果资料的取得是来自简单随机抽样（simple random sampling）的结果，此时数据不应该也不会有相关的。因此，在进行回归分析后，残差分析中残差项间不应该具有非独立的特性。也就是说，通过对残差图（residual plot）的检查，应该找不到残差项间的规则变化；以 Durbin Watson 值检验时，统计量 DW 值应该在 2 附近。有关于统计理论模型误差项的独立性，如果我们的数据很确定来自简单随机抽样的结果，则这个特性是可以达到的。

事实上，在社会科学或是管理学的横断面研究上，我们很难在抽样设计时做到完全简单随机抽样。基于现实的考虑，人们往往会将问卷发放到许多单位、组织或班级来填写，因此所搜集到的数据在某个程度上是有相关的。例如，最常看到的教育与心理研究，往往抽样设计是多阶段随机集群抽样（multi-stage random cluster sampling），是从许多的学校中抽取部分学校，再从所抽中的学校里抽出某一个班级或数个班级，然后对该班级进行全面问卷调查。假设我们想研究家庭社会经济状况对学生成绩的影响，我们所搜集的资料是每个学生父母职业与教育程度，以及该学生的成绩。此时，若将所有学校所搜集的学生与其父母资料汇集在一起，数据间的独立特性可能就不复存在了。所有的样本数据中，部分学生的数据彼此间会有关系，因为他们来自相同的学校班级，甚至来自相同的学区。因此，可能受到班级导师、学校校长教育理念的影响或是学区的因素甚至城乡差距都会关系到家庭社会经济状况对学生成绩的影响。在我们的回归模型设定里，并未考虑到除了学生数据以外的因素，而这些因素就会通通归为残差项。这时候，在这样的架构下，分析模型遗漏了重要的变量，或忽略了要控制的因素，导致残差项不再具有独立的特性，严重违反了回归模型的理论假设。

同样的问题也常发生在管理的组织研究上，我们经常将问卷发放到不同的组织中，请求各公司组织的员工或管理者填答，协助研究的进行。我们所搜集到的数据有可能来自相同公司的不同员工。此时若研究组织领导气氛对员工业绩的影响，我们会发现每家公司的领导气氛应该是一样的。但是员工的业绩会同时受到该公司组织领导气氛的影响，产生组织与员工分析层次的不同。在这样的研究中，组织特征是属于总体（macro）变量，而员工特征则是属于（micro）变量，因为每一家公司的每个员工都受到同一组织领导气氛的影响，即受到共同组织的领导

气氛这个情境或脉络（context）的影响，产生了共享的经验。除了个人因素外，组织的因素也会对员工业绩产生影响。当遗漏变量进行回归分析，将会得到残差项不独立的结果。这样资料搜集的方法，是属于一种嵌套、镶嵌或内属（nested）设计，这样的数据若利用一般的回归分析方法，残差项将违反模型独立性与同质性的假设。

1.2 遗漏重要自变量的后果

不管是随机集群抽样或是嵌套设计所搜集的数据，以一般的回归模型进行分析，都会出现忽略集群与嵌套内受试者间“相关”这一变量。也就是说，从一般线性模型的角度来看，就是回归模型遗漏了重要的自变量，就是集群或嵌套的关系，或称为“情境变量”或“脉络变量”（contextual variable）。在回归分析中，遗漏重要的自变量将会导致所估计出来的误差项方差被高估，所估计的回归系数标准误被低估的情况发生，使得在回归系数的假设检验里，零假设容易被拒绝，造成第一类错误（type I error）增大的问题。即，原本是不应该拒绝的零假设，但是我们却得到了统计量显著的结果，做出零假设被拒绝的结论。

当遗漏重要自变量与未遗漏重要自变量的两个回归方程相比较时，未遗漏重要变量的回归模型所估计出来的回归系数，要比遗漏重要自变量方程的估计值准确。但是遗漏重要自变量的方程，其回归系数的估计标准误却要比未遗漏重要自变量方程的回归系数估计标准误小，容易拒绝零假设，较易获得所欲得到的结论。至于一般线性模型中，存在许多自变量时，遗漏重要自变量对于回归分析的回归系数估计量的变异证明，可参阅计量经济学专著[1]。

增加无解释能力的自变量回归方程，所估计的回归系数不受影响，其估计值仍为总体参数的无偏估计，但是在估计上却因为多搜集变量数据造成效率降低。

1.3 忽略独立性的后果

在上一节中，讨论了遗漏重要自变量对于回归分析中回归系数估计量变异误的影响，这个重要自变量就是集群抽样，或是镶嵌设计中各个集群的特性或情境变量，也就是班级或公司组织特征。这些特征对学生成绩或者员工业绩会产生实

质影响。忽略这些变量当然在回归分析上会产生上述残差项独立性的问题，即使搜集到这些情境变量，也无法直接套入回归分析的模型中，因为这些班级、学校或是公司组织的特性放入回归模型中，正是需要分层线性模型（Hierarchical Linear Modeling）来克服上述问题。

在遗漏重要自变量的回归模型中，我们看到的是所要研究的变量 X_i 对 Y 的影响，其影响了除了回归系数估计量的无偏性和标准误的大小外，忽略了班级或组织内同一群学生和同一群员工受到班级与组织的影响，这种影响在组织研究中称为共享的经验，也会造成回归模型残差项产生相关，违反回归分析的误差项必须为独立的假设。当忽略集群抽样或镶嵌设计的数据间相关性、相似性的问题时，因为这个同一组织或同一学校或班级内受试者彼此之间相关性、相似性特征会被灌注到残差项里去，导致残差项不再具有独立性的特征，所以其不再具有独立性的特征将导致残差项方差的增加，也即现有的自变量无法完全充分解释因变量的变异。

总之，在随机集群抽样下，或是嵌套设计的情况下，对所搜集到的数据利用一般的线性模型进行分析，则原先模型设定的假设 $\varepsilon_i \sim iidN(0,\sigma^2)$ 将可能被违反。因为集群抽样的结果导致每个集群下的观测值在集群特征影响下都会具有某种共同的特质。例如，班级内的学生会受到教师的特性影响，或是同一部门的员工受到组织文化的影响。因而学生的成绩或是员工业绩表现上的差异，部分来自学生与员工个体层次属性所影响外，部分亦受到班级与组织的总体层次特质所影响。

在集群抽样与嵌套设计下，每一个观测值理论上彼此之间存在着某种程度的相似性，因此每个观测值所提供的信息不再是独立的，因而造成个体层次所提供的信息减少，不像在完全简单随机抽样中，每个个体都是独立的信息提供者。在集群抽样与嵌套设计的性质下，个体观测值所提供的信息会减少。这种信息的减少，是因为来自集群抽样下的相似性所造成的，Smith（2004）称之为设计效应（design effect）；若是来自于嵌套结构的相似性所致，则称之为相同效应（same effect）。将因变量的变异分解为“组织内变异”和“组织间变异”、纠正标准误差是多层模型技术主要的特点和优势。利用该模型，可以分析组织特征对个体的影响，探讨因变量的变异、自变量对因变量的影响如何因组织而异。